Data Access Technology Workshop
Position Paper
Thomas H. Hinke

Introduction

This paper addresses the seven questions that were requested of attendees prior to attending the Data Access Technology Workshop that is scheduled for October 8-9, 2002. Each of the questions is addressed in the following seven sections.

Question 1: What emerging technologies do you feel would best enhance user access to EOSDIS data and why?

Grids are an emerging technology that can take EODSIS to the next level that would permit users to access data from the EOSDIS archives and move it to one or more grid-resident processors, where their applications could process the data and deliver the results. Under this vision, the grid would provide a set of tools that would allow users to easily specify the flow of data from one or more EOSDIS archives, through one or more processing and storage services, with the result being some processing result desired by the user.

Emerging grid technologies, such as NASA's Information Power Grid (IPG), provide a set of tools and systems that provide the following capabilities:

1. An underlying security infrastructure, so that authentication information (and other information if desired) is protected from unauthorized access.
2. A single sign-on capability that allows a user to identify and authenticate himself once per session, with the grid supporting his use of any grid resources to which he has been permitted use, with no additional authentication required.
3. A set of tools for securely moving data between multiple, distributed grid resources.
4. A set of tools to permit users to securely run applications on multiple, distributed grid resource.
5. A grid information service that permits grid users to discover needed information about grid resources.
6. An archival storage system (such as the Storage Resource Broker, developed at the San Diego Super Computer Center) that is grid-enabled , supports the storage of data on both secondary and tertiary storage, and supports a metadata catalog to facilitate discovery of desired data.
7. A job broker that assists users in identifying appropriate grid resources that are appropriate for a running a particular application.
8. A job manager to manage a set of jobs that the user might need to run in support of an application.

One of the latest thrusts of the grid community is to fuse the emerging web services technology with grid technology in order to provide grid-enabled web services. Grid-enabled web services are grid-accessible -applications whose interface is described by an XML(Extensible Markup Language) based interface, accessed via a standard protocol such as SOAP (Simple Object Access Protocol), with security provided by the underlying grid security infrastructure. Grid-enabled web services are designed to be easily included as components of more complex services, much as objects in an object-oriented programming language can be reused and combined into more complex objects. The goal is to have a set of reusable services that can provide the building blocks for various application-specific services accessible on the grid.

Question 2: How do you envision these technologies being used? How should they be used in combination with each other?

One vision is that of a scientist (or other user of EOSDIS data) using grid tools to select data to be processed and then designing a processing plan that specifies the sequence of grid-based processing services that are to be used to achieve the desired results. The scientist can then use a grid broker to determine the best (e.g., cheapest, fastest) set of grid services that can be used to support the execution of this processing plan. This set of grid services could include some processing at the user's computer, as well as more extensive processing using resources and services available on the grid. The task would then be launched on the grid and controlled by workflow management software that ensures that the task is completed and the results delivered.

Under this vision, the grid could provide services that would permit the user to easily tap into archived data, easily perform transformations that would make this data more usable to the user or his application (e.g., sub-setting or reformatting data), and then support user specified processing.

In addition, grid services that support the Open GIS Consortium standards could be made available on the grid. The grid could provide the infrastructure that would permit users to conveniently extract data from an archive and move it to various Open GIS based services that were available on the grid. In this way, the user could utilize the services of the grid for those capabilities that were not available on his local workstation or computer, or that required processing that was of a magnitude greater than one the user had locally.

Grids could also be used for the ongoing work of EOSDIS product generation, with the various archives being able to load-share processing among product generation centers.

In the future, it might be desirable to establish a grid in space with the ability for users to establish processing sequences that involve both data from Earth-based archives as well as real-time data flowing from instruments in orbit, since one of the goals of grid technology is to integrate instruments, with computational and data resources.

Question 3: Can the technologies be readily leveraged to achieve these results today? If not, when do you think they will be viable?

Existing NASA grid work can be readily leveraged to support the proposed vision. The information power grid is NASA's implementation of a grid. In its current state, it has 24x7 support provided as part of support that the NASA Advanced Supercomputing (NAS) Division provides at the NASA Ames research center for its production supercomputers. The IPG is available today on a number of computers currently spread over three NASA centers, with additional grid work being complemented at three others.

The IPG projects helps support and is constructed on top of Globus software, which supports items 1 through 5 under question 1. The Globus software is being developed by Argonne National Laboratory, the University of Southern California's Information Sciences Institute, the National Center for Supercomputing Applications and the San Diego Supercomputer Center. While not the only grid software in the world, for much of the world, the Globus software suite has become a de facto standard. A number of companies are planning on offering commercial versions of this software.

To support access to files, the Storage Resource Broker (developed by the San Diego Supercomputer Center) provides a grid-enabled, distributed file management system that has been used for a number of applications. This supports item 6 under question 1.

The remainder of the of the items (7 and 8 under question 1) are being supported under the NASA Computer Information Communications Technology (CICT) program that will be taking the IPG to a reference implementation for various NASA enterprises by December 2004. Current work includes the development of a broker and a job manager, which have been successfully used in satisfying a recent program milestone that involved use of the grid by a number of grand-challenge applications under the CICT program. The IPG project is deploying a pre-production grid that is currently being used by various applications and developing software that is intended to make the grid easy to install, easy to maintain and easy to use. The goal is a less-than eight-hour installation and automated tools to help maintain the health of the grid. Work to facilitate the use of the grid is continuing under the IPG project as well as other grid projects in Europe, the UK and an emerging grid work in Asia.

Question 4: What investments do you think EOSDIS should make in these technologies today? Over the next 5 years?

EODSIS should make investments to either connect a subset of existing archives to a grid or at least develop a pilot program in which a number of mirror archives are connected to the grid. These connected archives should include a sufficiently useful slice of Earth science data that scientists or other EOSDIS data users will be motivated to use this data.

EOSDIS should also make investments in sufficient grid processing resources such that scientists will be motivated to use the grid. One approach is to work with the IPG in order

to supplement the existing processing capabilities currently available through the IPG and work to make these resources available to scientists that are interested in using the grid.

EODSIS should also make investments that will motivate some early-adopter users of EOSDIS data to use the grid for their processing. The CICT program has done that through the identification of some grand-challenge applications that are being developed to use the IPG.

The above investments could be made in the relatively near future to use the existing grid capabilities that are already there and to grid-enable some EOSDIS data archives.

Over a five year period, EOSDIS should make investments to provide grid services that are needed by a large set of Earth scientists or other EOSDIS data users and to provide the processing capabilities needed to support these services in a production system. EOSDIS should also make any investments in grid software necessary to more efficiently support Earth science applications. These would be issues that were identified as part of the near-term pilot efforts.

Question 5: What are reasonable cost expectations for leveraging these technologies as you've suggested?

This depends upon how this is approached. One could grid-enable one or more existing EOSDIS archives or one could establish one or more pilot archives that contain a useful slice of data from one or more of the existing EOSDIS archives. In the first option, the cost would be relatively low to connect the archive to the grid. One approach would be to use the SRB as the grid-enabled file system interface, configure the SRB to have access to some or all of the archives data files, install the IPG software on the SRB processor and then ensure that a high-speed link is provided. The SRB provides the ability to mediate access to the data, so access to this grid-enabled portion of the archive could be controlled, although one could still run into congestion if too many users attempted to access the system at the same time.

The other possibility is to develop a mirrored version of a subset of one or more archives, using the SRB to provide a pilot grid-enabled archive that would be separate from any existing archive, so as not to interfere with its production work. This would require the purchase of some additional processors and disks for each archive. With the cost of hardware and disks dropping, this could be less than a hundred thousand dollars for each pilot archive. To be most effective for investigating the applicable technologies, the archive should also provide access to data stored on tertiary storage, and this could increase cost somewhat if mass storage systems had to be provided for each pilot archive. One approach would be to use an existing IPG mass storage system that is currently being considered for upgrade. In this case, the cost could be fairly small to provide some additional tape drives and tapes. Each archive would have to have the grid software installed and would have to be connected to the grid via some reasonably high-speed network connections.

Finally, a set of processing resources would have to be provided, perhaps by supplementing those already available through the IPG or by connecting some existing Earth science processing resources to the grid.

Depending upon the number of users to be supported, the costs for the above could range from a few hundred thousand dollars to a couple of million.

Question 6: What are the anticipated impacts on end users if these technologies are deployed, including investments required?

For application developers, the grid can lead to more rapid development of services since it provides those common functions that any distributed application would have to develop. Since the grid has already provided these capabilities, the developers of these applications will not have to re-implement them, but will be able to build on these already existing services. Experience shows that the existence of these grid services greatly increases the ease of developing distributed applications.

In addition, the grid has spawned an international community of grid development work in Europe, the United Kingdom , North America and Europe. These efforts are all participating in the Global Grid Forum, which seeks to develop best practices and standards for grids. The Global Grid Forum is structured after the IETF (Internet Engineering Task Force), which has been the international organization that has developed internet standards. The Global Grid Forum meets three times a year as a collection of research and working groups, each working to facilitate the development of grid technology.

This has spawned commercial interest in grids with a number of companies committing to the development of commercial grid offerings. The last Global Grid Forum meeting in Edinburgh, Scotland in July 2002 had approximately 900 people in attendance, which was a 500 person increase over the previous meeting in Toronto, Canada in February 2002. This indicates a growing international interest in grid projects, which now are very active in Europe, the United Kingdom and various places in Asia including Korea, Japan and China.

Because of this growing world-wide work in grid technology, EOSDIS will be able to take advantage of grid developments that may occur in various parts of the world. Most of the grid projects are interested in sharing grid developments among the various grid development efforts.

For the users of these grid services, their gain is the more rapid development of new services that can help them more easily use EOSDIS data. While their particular application may hide the underlying grid from the users, they will benefit from the grid's support for the grid-enabled application that they are using. If the grid provides a set of commonly used Earth science data services and if the EOSDIS archives are connected to the grid, then users should be able to more rapidly process their data and gain needed

results. This should speed up their work and lead to more useful results coming from EODSIS data.

The Earth science community should consider investing in, and working with the grid community to provide the services required for Earth science applications. In addition, the Earth science community should invest in pilot systems that provide grid-enabled access to a subset of EOSDIS data and should work to encourage early-adopter scientists to begin to experiment with using grid technology.

This could take the form of funding to either develop an Earth science grid or to become part of the Information Power Grid. The minimum investment would be to provide the ability to connect the archives to a grid such as the Information Power Grid and some support for the provisioning of necessary computational resources that would need to be available on the grid. The investment could range from a few hundred thousand dollars up to millions of dollars, depending upon the nature of the new computational resources to be provided.

Another EOSDIS thrust could provide funding for research personnel to become early adopters of grid technology, with the funding supporting the transformation of existing applications to the grid or the development of new applications. The NASA CICT program has taken this approach by identifying a number of grand challenge applications that are funded to perform their research by using the Information Power Grid. The CICT program current includes some Earth Science requirements that are anticipated to fund grid work that is supportive of Earth Science needs and will complement funded programs coming from the Earth Science community that will use the grid capabilities provided. The Earth science community should consider working with the CICT program to ensure that Earth science funded activities and CICT funded activities are complementary.

Question 7: What unique researcher data access requirements need to be supported?

Researches using the grid will need to be able to access data from both secondary and tertiary storage. Tertiary storage access poses challenges not associated with secondary storage access, due to the time it takes to retrieve data from secondary storage. The tertiary storage system's robotic arm must retrieve the desired tape, mount it on a drive and then transfer data to disk. This can take on the order of minutes, versus milliseconds for access to secondary storage. Since there are a limited number of drives and each drive could be occupied for a number of minutes to retrieve a single tape, there could be a considerable amount of contention for access to grid-accessible mass storage systems, which will need to be scheduled so that users are granted access on some acceptable priority.

To help mitigate this problem, research needs to be performed on caching schemes that will permit data to be cached on secondary storage so that the need to actually retrieve data from tertiary storage is reduced.

Another area of possible research is providing capabilities to control the flow of data from the archive to the grid resource where the processing is occurring. The concern is that these system may not have sufficient memory or disk space to handle all of the data required for a particular processing effort (e.g., data mining), thus the challenge is to ensure that the data gets there when needed, but not too much prior to need, so that data does not overrun the site. Of course, this may be mitigated as the cost of disk storage and processor memory drops, if grid resources have sufficient storage capability to support the desired processing.